ICDAR2013 Competition on Historical Book Recognition – HBR2013[†]

A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher
Pattern Recognition and Image Analysis (PRImA) Research Lab
School of Computing, Science and Engineering, University of Salford
Greater Manchester, M5 4WT, United Kingdom
www.primaresearch.org

***Abstract*—This paper presents an objective comparative evaluation of layout analysis and recognition methods for scanned historical books. It describes the competition (modus operandi, dataset and evaluation methodology) held in the context of ICDAR2013 and the 2nd International Workshop on Historical Document Imaging and Processing (HIP2013), presenting the results of the evaluation of five methods - three submitted and two state-of-the-art systems (one commercial and one open-source). Three scenarios are reported in this paper, one evaluating the ability of methods to accurately segment regions, one evaluating segmentation and region classification (with a text extraction goal) and the other the whole pipeline including recognition. The results indicate that there is a convergence to a certain methodology, in terms of layout analysis, with some variations in the approach. However, there is still a considerable need to develop robust methods that deal with the idiosyncrasies of historical books, especially for OCR.**

Keywords - performance evaluation; page segmentation; region classification; layout analysis; recognition; OCR; datasets; historical documents;

I. INTRODUCTION

Vast repositories of historical books are being scanned or plans exist for them to be scanned in libraries and archives around the world. There is a very significant need for full-text extraction and recognition systems that can be used in such large-scale digitisation projects [1]. Currently, if OCR is applied to the scans, its results have relatively low word accuracy. There is relatively little choice in terms of OCR systems and their capabilities to be trained for historical fonts and to use historical dictionaries [1][2][3].

There are distinct steps in the recognition workflow whose performance is crucial to the overall success of the system. First, Layout Analysis (comprising Page Segmentation and Region Classification) is one of the most well-researched fields in Document Image Analysis, yet new methods continue to be reported in the literature, indicating that the problem is far from being solved. Frequently, methods are devised with a specific application in mind and are fine-tuned to the image dataset used by their authors. However, the variety of documents encountered in real-life situations (and the issues they raise) is far wider than the target document types of most methods.

Second, (machine-printed) OCR which has been largely abandoned by academic researchers, is encountering severe challenges in historical documents due to archaic fonts, degraded quality of glyphs and inapplicability of modern lexica [1], among others. Systematic evaluation is crucial to study the issues and attempt to make progress.

The aim of the ICDAR Page Segmentation competitions (since 2001) has been to provide an objective evaluation of methods, on a realistic contemporary dataset, enabling the creation of a baseline for understanding the behaviour of different approaches in different circumstances. This is the only international layout analysis competition series that the authors are aware of. Other evaluations of page segmentation methods have been constrained by their use of indirect evaluation (e.g. the OCR-based approach of UNLV [4]) and/or the limited scope of the dataset (e.g. the structured documents used in [5]. In addition, a characteristic of previous reports has been the use of rather basic evaluation metrics. While the latter point is also true to some extent of early editions of this competition series, which used precision/recall type of metrics, the 5th edition of the ICDAR Page Segmentation competition (ICDAR2009) [6] made significant additions and enhancements. First, that competition marked a radical departure from the previous evaluation methodology. A new evaluation scheme was introduced, allowing for higher level goal-oriented evaluation and much more detailed region comparison. In addition, the datasets used since then have been selected from new datasets [7][8] that contain different instances of realistic documents.

This edition (HBR2013) is based on the same principles established by the 2011 competition on historical document layout analysis [8] but its focus is on the *complete recognition workflow* for books, reflecting the significant need to identify robust and accurate methods for large-scale digitisation initiatives. HBR2013 is co-sponsored by ICDAR2013 and HIP2013 (2nd International Workshop on Historical Document Imaging and Processing).

An overview of the competition and its modus operandi is given next. In Section 3, the evaluation dataset used and its general context are described. The performance evaluation method and metrics are described in Section 4, while each of the participating methods is summarised in Section 5. Finally, different comparative views of the results of the competition are presented and the paper is concluded in Sections 6 and 7, respectively.

[†] This work has been funded through the EU 7th Framework Programme grant SUCCEED (Ref. 600555)

1520-5363/13 $26.00 © 2013 IEEE
DOI 10.1109/ICDAR.2013.294

II. THE COMPETITION

HBR2013 had the following three objectives. The first was a comparative evaluation of the participating methods on a representative dataset (i.e. one that reflects the issues and their distribution across library collections that are likely to be scanned). Delving deeper, the second objective was a detailed analysis of the performance of each method in different scenarios from the simple ability to correctly identify and label regions to a text recognition scenario where the reading order needs to be preserved. This analysis facilitates a better understanding of the behaviour of methods in different digitisation scenarios across the variety of documents in the dataset. Finally, the third objective was a placement of the participating methods into context by comparing them to leading commercial and open-source systems currently used in industry and academia.

The competition proceeded as follows. The authors of candidate methods registered their interest in the competition and downloaded the *example* dataset (document images and associated ground truth). The *Aletheia* [10] ground-truthing system (which can also be used as a viewer for results) and code for outputting results in the required PAGE format [11] (see below) were also available for download. Three weeks before the competition closing date, registered authors of candidate methods were able to download the document *images* of the *evaluation* dataset. At the closing date, the organisers received both the executables and the results of the candidate methods on the evaluation dataset, submitted by their authors in the PAGE format. The organisers then verified the submitted results and evaluated them.

III. THE DATASET

The importance of the availability of realistic datasets for meaningful performance evaluation has been repeatedly discussed and the authors have addressed the issue for contemporary documents by creating a dataset with ground truth [4] and making it available to all researchers. In comparison, representative datasets of historical documents are even more difficult to collect (from different libraries) and to ground truth (due to the nature and variety of the texts).

Under the direction of the authors a comprehensive dataset of historical document images has been created as part of the IMPACT project [1] and is now available through the IMPACT Centre of Competence in Digitisation [12]. The dataset contains approximately 700,000 page images (with associated metadata) from 15 different content holders, including most national and major libraries in Europe. This dataset has been collected to not only reflect the conditions and artefacts of historical documents that affect document analysis, but also the needs and priorities of the libraries, in terms of what types of documents (representative of their holdings) dominate their digitisation plans. The complete dataset consists of printed documents of various types, such as books (approximately 355,000 pages), newspapers, journals and legal documents, in 25 different languages and 11 scripts, from the 17th to the early 20th century.

The unique value of this dataset though is significantly enhanced by the availability of a considerable volume of detailed ground truth. In total, 52,000 images (42,000 book pages) have been ground truthed at the level of regions (equivalent to paragraphs, illustrations, separators etc.). In addition to the accurate description of region outlines, the text contained in each (textual) region has been re-keyed under strict rules, preserving typographic conventions, including, abbreviations, ligatures etc.

Figure 1. Sample evaluation set images (not shown to scale).

For the purpose of this competition, 100 book page images were selected from the IMPACT dataset as a representative sample from different ages ensuring the presence of different issues affecting layout analysis and OCR. Such issues include dense printing (minimal spacing), irregular spacing, varying text column widths, presence of separators, marginal notes and a variety of languages (English, French, German and Spanish) in both Latin and Fraktur scripts. Sample pages can be seen in Fig. 1.

It is worth noting that the images for this competition were selected so as to be as realistic as possible, in some cases suffering from moderate bleed-through, page curl, containing image borders etc.

The ground truth is stored in the XML format which is part of the PAGE (Page Analysis and Ground truth Elements) representation framework [11]. For each region on the page there is a description of its outline in the form of a closely fitting polygon. A range of metadata is recorded for each different type of region. For example, text regions hold information about *language*, *font*, *reading direction*, *text colour*, *background colour*, *logical label* (e.g. heading, paragraph, caption, footer, etc.) among others. Moreover, the format offers sophisticated means for expressing reading order and more complex relations between regions. Sample images with ground truth description can be seen in Fig. 2.

Figure 2. Sample images showing the region outlines (blue: text, magenta: separator, green: graphic, cyan: image).

IV. Performance Evaluation

A. Layout Analysis

The performance analysis method used for this competition can be divided into three parts. First, all regions (polygonal representations of ground truth and method results for a given image) are transformed into an interval representation [9], which allows efficient comparison and calculation of overlapping/missed parts. Second, correspondences between ground truth and segmentation result regions are determined. Finally, errors are identified, quantified and qualified in the context of one or more application scenarios.

The region correspondence determination step identifies geometric overlaps between ground truth and segmentation result regions. In terms of Page Segmentation, the following situations can be determined:

- *Merger*: A segmentation result region overlaps more than one ground truth region.
- *Split*: A ground truth region is overlapped by more than one segmentation result region.
- *Miss (or partial miss)*: A ground truth region is not (or not completely) overlapped by a segmentation result region.
- *False detection*: A segmentation result region does not overlap any ground truth region.

In terms of Region Classification, considering also the *type* of a region, an additional situation can be determined:

- *Misclassification*: A ground truth region is overlapped by a result region of another type.

Based on the above, the segmentation and classification errors are *quantified*. This step can also be described as the collection of raw evaluation data. The amount (based on overlap area) of each single error is recorded.

Having this raw data, the errors are then *qualified* by their significance. There are two levels of error significance. The first is the implicit *context-dependent* significance. It represents the logical and geometric relation between regions. Examples are *allowable* and *non-allowable* mergers. A merger of two vertically adjacent paragraphs in a given column of text can be regarded as allowable, as the result of applying OCR on the merged region will not violate the reading order. On the contrary, a merger between two paragraphs across two different columns of text is regarded as non-allowable, because the reading order will be violated in the OCR result. To determine the allowable/non-allowable situations accurately, the reading order, the relative position of regions, and the reading direction and orientation are taken into account.

The second level of error significance reflects the additional importance of particular errors according to the application scenario for which the evaluation is intended. For instance, to build the table of contents for a print-on demand facsimile edition of a book, the correct segmentation and classification of page numbers and headings is very important (e.g. a merger between those regions and other text should be penalised more heavily).

Both levels of error significance are expressed by a set of weights, referred to as an *evaluation profile* [9]. For each application scenario to be evaluated there will be a corresponding evaluation profile.

Appropriately, the errors are also weighted by the size of the area affected (excluding background pixels). In this way, a missed region corresponding to a few characters will have less influence on the overall result than a miss of a whole paragraph, for instance.

For comparative evaluation, the weighted errors are combined to calculate overall error and success rates. A non-linear function is used in this calculation in order to better highlight contrast between methods and to allow an open scale (due to the nature of the errors and weighting).

B. Text Recognition

For the evaluation of OCR results a word-based method has been implemented. The order of the words is not considered (Bag of Words) since the reading order of the submitted results is not known and a manual serialization of the text is too cumbersome.

Words for both ground truth and OCR result are extracted separately in two steps. First, the text content of each region is separated into words using white spaces and punctuations. Second, the text is integrated into a look-up table with "Word" and "Count" as columns.

The two resulting tables are then compared by identifying missed words and falsely detected words. The Success Rate measure defined takes into account the correct count of words (i.e. how many of the instances of each word on a page have been correctly recognised)

V. Participating Methods

Brief descriptions of the methods whose results were submitted to the competition are given next. Each account has been provided by the method's authors and edited (summarised) by the competition organisers.

A. The EPITA method

This method [13] was submitted by Guillaume Lazzara, Roland Levillain, Thierry Géraud, Yann Jacquelet, and Julien Marquegnies of EPITA, France. It is a bottom-up approach based on connected-component aggregation. First, the document is binarised using a multiscale implementation of Sauvola's algorithm. Vertical and horizontal separators are then identified, removed and the document is denoised.

The remaining components are labeled and from those similar component groups, component alignments and white spaces (on their sides) are determined. These virtual delimiters associated with separators provide a structure of the different blocks in the document. Using this information, component groups are merged to create text lines.

Subsequently, lines are linked into text regions. Text indentations, spaces between adjacent lines and text line features are then analysed in order to split regions into paragraphs. Paragraphs overlapping significantly are also merged together. Among the part of the documents where no text has been found, the components are retrieved and considered as images. Finally, some cleanup is performed: separators detected in images, in paragraph and in document borders are removed, false positive text areas are removed in images and borders and small images included in text areas are considered as drop capitals.

This is the same method as submitted to the ICDAR2011 competition [8]. It is developed using the SCRIBO module [14] and the source code is freely available [15].

B. The Jouve method

This method was submitted by Michaël Fontaine and Mohamed Zayed of JOUVE, France [16], a commercial organisation specializing in digitisation services.

The Layout Analysis subsystem is essentially the same as the one submitted to (and won) the 2011 Historical Document Layout Analysis competition [8]. The main principle of the method is to identify and extract regions of text by analysing connected components constrained by black and white (background) separators – the rest is filtered out as non-text. First, the image is binarised, any skew is corrected and black page borders are removed. Subsequently, connected components are extracted and filtered according to size (very small components are filtered out). By analysing the size and spacing of the components (using global and local information), characters and words are identified. Black horizontal and vertical lines (corresponding to separators) are also identified in the size filtering step. White separators corresponding to space between columns are then identified by aggregating white rectangles aligned at the end of words and filtering out non-viable separators. With the aid of white separators, words are grouped into text lines without risking merging words belonging to different columns.

Text lines of the same height and located at the same distances are grouped to reconstitute the paragraphs. Paragraphs are finally merged in order to obtain columns guided by both the black and the white separators detected. The reading order is determined by an iterative method using vertical white streams, horizontal and vertical black separators, and a heuristic to sort boxes.

For character recognition, JOUVE takes the approach of font-training per book. For each book, character clustering is done in order to obtain a set of similar characters. The method used to build the cluster is the one used in Leptonica library for the JBIG2 compression. Images of the representative character of the clusters are labelled by human operators and all the characters of the clusters are automatically labelled by propagating the label of their respective representative character. These clusters are used in order to train a Recurrent Neural Network. For HBR2013 (containing pages from a variety of books and different fonts) the mixture of all learnt characters shapes coming from the whole *evaluation* dataset were used.

C. The PAL method

This bottom-up approach focuses on extracting the regions of text from the image, ignoring non-text regions (based on [17]). It was submitted by Kai Chen, Fei Yin and Cheng-Lin Liu of the National Laboratory of Pattern Recognition (NLPR) at the Institute of Automation of the Chinese Academy of Sciences. After the image is binarised, the method starts by extracting the foreground connected components (CC). Then whitespace between vertically adjacent CCs is extracted in the form of small rectangles, referred to as horizontal cut rectangles (HCR). Subsequently, horizontally adjacent HCRs are linked into chains (HCRC), which are used to identify horizontally adjacent CCs. CC chains (CCC) are then formed by linking horizontally adjacent CCs. After examining the gaps between neighboring CCs inside each CC chain, the chain is cut into initial text lines where the gaps are relatively wider. Whitespace between horizontally adjacent initial text lines is also extracted in the form of small rectangles, referred to as vertical cut rectangles (VCR). For each short initial text line which has two VCRs at both ends, the narrower VCR is eliminated and the two text lines which are horizontally adjacent to this VCR are merged to form a new text line. A VCR can also be eliminated if it is surrounded by text lines in four directions (above, below, left and right), and the left and right neighboring text lines are merged. The remaining VCRs are clustered into groups by linking vertically adjacent ones. Each group is analysed by comparing it with the already eliminated VCRs. If the difference is not obvious, the whole group is eliminated and text lines involved are merged. Finally, the vertically adjacent text lines are linked into text blocks.

VI. RESULTS

Evaluation results for the above methods are presented in this section in the form of graphs with corresponding tables. For comparison purposes, the layout analysis and recognition components of a leading product, ABBYY FineReader® Engine 10 (FRE10), and that of the popular open-source system, Tesseract 3 are also included. It must be noted that both FRE10 and Tesseract 3 have been evaluated out of the box, with no training or knowledge of the dataset.

Three scenarios have been defined for the competition - two layout evaluation profiles plus performance of OCR. The first profile is used to measure the pure segmentation performance. Therefore, misclassification errors are ignored completely. Miss and partial miss errors are considered worst and have the highest weights. The weights for merge and split errors are set to 50%, whereas false detection, as the least important error type, has a weight of only 10%. Results for this profile are shown in Fig. 3.

The second profile is basically equal to the first one except that it also includes misclassification. As the main fo-

cus lies on text, misclassification of text is weighted highest. All other misclassification weights are set to 10%. Results for this profile are shown in Fig. 4. A breakdown of the layout analysis errors made by each method is given in Fig. 5.

Finally, the OCR performance of the only submitted method (JOUVE) that includes recognition is compared with the state-of-the-art systems in Fig. 6.

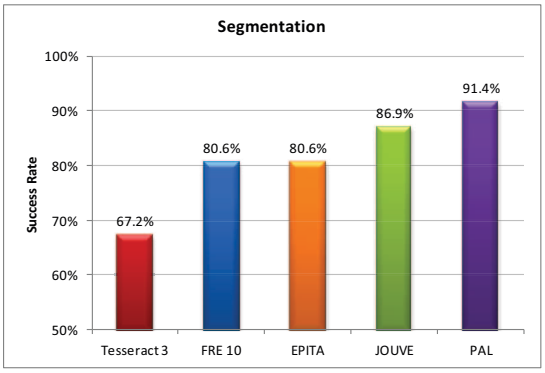

Figure 3. Results using the segmentation evaluation profile.

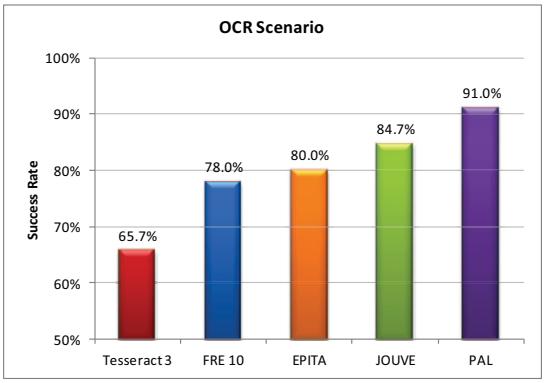

Figure 4. Results using the OCR-scenario evaluation profile.

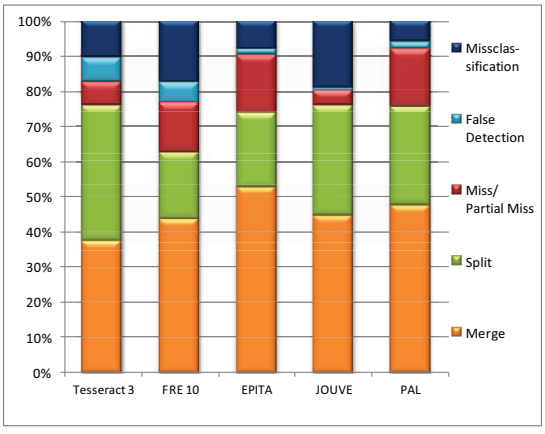

Figure 5. Breakdown of errors made by each method.

VII. Concluding Remarks

The three systems follow a similar bottom-up layout analysis approach but their performance varies on book images. In terms of recognition, JOUVE may achieve better performance if trained and applied on specific books but the state-of-the-art systems seem more flexible. The lower relative performance of Tesseract is mostly due to worse image enhancement and overlapping region descriptions. The results show that the PAL method has an overall advantage, especially in the OCR scenario. It is clear that there is still a considerable need to develop robust methods that deal with the idiosyncrasies of historical books.

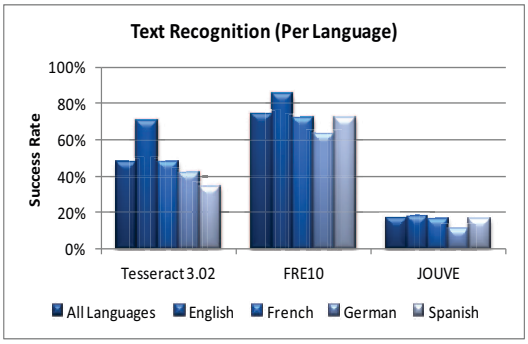

Figure 6. Recognition performance per language.

References

[1] IMPACT project: http://www.impact-project.eu

[2] Europeana Newspapers project: http://www.europeana-newspapers.eu

[3] The Early Modern OCR Project, http://emop.tamu.edu

[4] J. Kanai, S.V. Rice, T.A. Nartker and G. Nagy, "Automated Evaluation of OCR Zoning", *IEEE PAMI,* 17(1), 1995, pp. 86-90.

[5] F. Shafait, D. Keysers and T.M. Breuel, "Performance Evaluation and Benchmarking of Six Page Segmentation Algorithms" *IEEE PAMI,* 30(6), 2008, pp. 941–954.

[6] A. Antonacopoulos, S. Pletschacher, D. Bridson, C. Papadopoulos, "ICDAR2009 Page Segmentation Competition", *Proc. ICDAR2009*, Barcelona, Spain, July 2009, pp. 1370-1374.

[7] A. Antonacopoulos, D. Bridson, C. Papadopoulos and S. Pletschacher, "A Realistic Dataset for Performance Evaluation of Document Layout Analysis", *Proc. ICDAR2009*, Barcelona, Spain, July 2009, pp. 296-300.

[8] A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher, "Historical Document Layout Analysis Competition", *Proc. ICDAR2011*, Beijing, China, Sept 2011.

[9] C. Clausner, S. Pletschacher and A. Antonacopoulos, "Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods", *Proc. ICDAR2011*, Beijing, China, Sept 2011.

[10] C. Clausner, S. Pletschacher and A. Antonacopoulos, "Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments", *Proc. ICDAR2011*, Beijing, China, 2011.

[11] S. Pletschacher and A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", Proc. ICPR2008, Istanbul, Turkey, August 23-26, 2010, IEEE-CS Press, pp. 257-260.

[12] IMPACT Centre of Competence: http://www.digitisation.eu

[13] Online demo: http://caroussel.lrde.epita.fr/olena/demos

[14] G. Lazzara, R Levillain, T Géraud, Yann Jacquelet, Julien Marquegnies, Arthur CrÈpin-Leblond, "The SCRIBO Module of the Olena Platform: a Free Software Framework for Document Image Analysis" *Proc. ICDAR2011*, Beijing, China, Sept 2011.

[15] Source available in the git repository: git://git.lrde.epita.fr/olena - branch: icdar/hnla2013 - location: scribo/src/contest/hnla-2013

[16] JOUVE – http://www.jouve.com

[17] Kai Chen, Fei Yin and Cheng-Lin Liu, "Hybrid Page Segmentation with Efficient Whitespace Rectangles Extraction and Grouping", Proc. ICDAR2013, Washington DC, USA, Aug 2013.